Analysis of Using Machine Learning for the Classification of Particles in Jets

Alan Tondryk, Sookhyun Lee

Abstract

Particle accelerators have helped humanity peer into the depths of the world's building blocks, from the most fundamental quarks and leptons to composite particles such as hadrons and mesons. A meson is made up of a quark and antiquark pair, while a hadron is made up of three quarks. At high energy particle accelerators, these leptons and composite particles are often generated in collisions forming a collimated spray which leaves signals in detectors. This pattern of particles is called a "jet". However, deciphering the particle contents of the jets that come from the particle collisions is not the easiest task. One particular issue is the reliable detection and identification of particles. At the sPHENIX experiment that will be built at RHIC located in NY on Long Island, jets are one of the main observables anticipated to improve our knowledge about the inner workings of quarks and gluons inside matter. As part of efforts towards developing software algorithms to reconstruct constituent particles inside a jet with high accuracy, machine learning techniques are being considered. This paper examines, as a first step, the improvement in the accuracy and purity of detections of photons resulting from the decay of neutral pions through the use of neural networks with the objective of classifying particles based on detections in the electromagnetic calorimeter. The possibility of extending this work to full software that uses information from all detectors is also discussed.

Introduction

Particle accelerators are used to find out what some of the fundamental physics that govern the world are. And as such, it is important that the corresponding particle detectors are able to characterize and identify particles coming out of the jets produced by particle collisions at particle accelerators. Reconstructing a list of particles from signatures such as charge, position, momentum, or energy in the detectors is dubbed “Particle Flow.”[1] Identification is done via an array of detectors nested around the point of collision of the particles. To properly characterize particles, different parts of the detector array are used for different things. For example, the tracking detector can track electrically charged particles to determine their momentum, the electromagnetic calorimeter measures the energy of particles (Such as photons and electrons) deposited in the detector as a result of electromagnetic interaction, and the hadronic calorimeter measures the energy of particles that go through nuclear interactions with the detector material such as charged pions, kaons and protons.[3] The intractability of these particles with detectors is shown in figure 1.

The focus here is in enhancing the particle identification capabilities of the current software on the sPHENIX detector.[4] One example is neutral pion decays. Neutral pions, meson particles with up or down type quarks, are commonly created during particle collisions, and most of the time these particles decay into two photons which fly out from the point of the collision towards the detector walls. Since photons interact electromagnetically, the energy would mostly be deposited in the electromagnetic calorimeter. Neutral pions tend to decay into two photons with an angle in between each other that depends on the momentum of a neutral pion. As momentum of a pion becomes higher, the decay photons hit really close to each other on the detector resulting in a merged cluster. Clusters are groups of energy detecting towers which show the spread out deposition of energy caused by the particle. In a merged cluster it is classically very hard to know from which particle the energy came, thus muddying the purity of detections and making it harder to reconstruct jets. Merged clusters are currently removed from data samples, based on the probability of the shower shape that is determined from analyzing test beam data.

We demonstrated as a proof of concept, improved purity and efficiency of detections in the case of merged clusters originating from neutral pions by using deep neural network based machine learning to identify jet constituents. The idea was to let the neural network find patterns in our data and be able to reconstruct the energy images into particles based on multiple classes of particles. Thanks to the network’s capability of learning complex dependency of various factors that determine the shape of a cluster, this approach let us classify particles with much greater ease and purity and also allowed us to use these merged clusters to our advantage instead of just throwing them out.

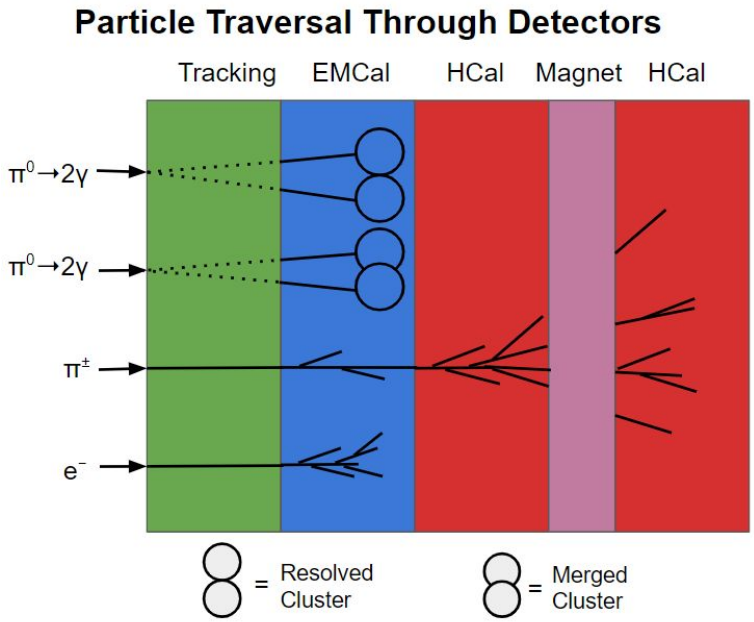

Figure 1) Different particles traverse different lengths of the detector structure effectively and interact in various parts.

Methods

Data Collection

Our method of data collection involved using a particle simulator. Using the built in framework for running particle simulations used by the phenix detector, we first ran just photons to study variables the program outputs and decide what variables we needed. Then, high transverse momentum neutral pions were generated and fed into the Pythia6 decayer that simulates the decay mechanisms for pions. The data samples obtained in this manner contained both resolved photon clusters as well as merged photon clusters which we were interested in classifying. Data such as energy, eta, phi, transverse momentum (pt), and shower profile was saved as these elements would be used later as features, a kind of input to the network. In addition, a two-dimensional distribution of energy in a cluster was saved in data, tagged as an image, a second kind of input to the network. Each image represents a cluster that showed a 5 pixel by 5 pixel map of where a particle hit. Each pixel in this case was a tower in the detector. We were able to visualize these images as heatmaps of energy in each tower. See figure 2a & 2b.

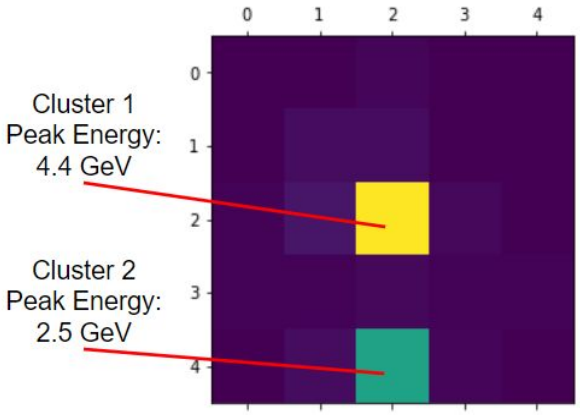

Figure 2a) EMCal Detector image for resolved cluster (class 0) caused by neutral pion to photon decays in the 15-20 GeV range. Brighter colors signify higher energies.

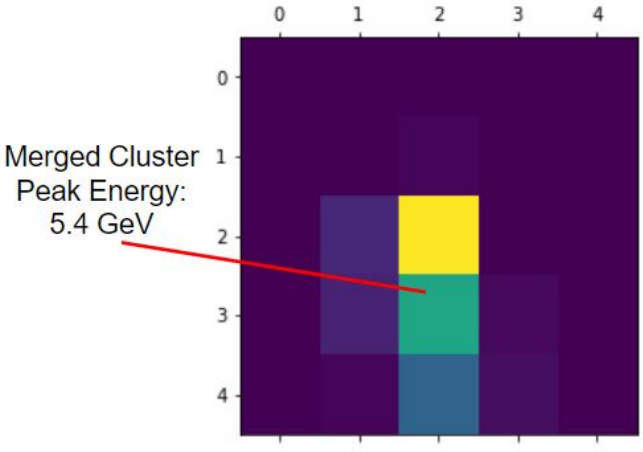

Figure 2b) EMCal Detector image for merged cluster (class 1) caused by neutral pion to photon decay in the 15-20 GeV range. Brighter colors signify higher energies.

Data Processing

Before processing our images with machine learning in a neural network, we had to format the data. Using numpy functions in python, the comma separated values (csv) data from the simulator containing our images was transformed into a tensor to comply with Keras, a machine learning tool. This was then able to be fed into our neural network along with the feature data we saved such as eta, phi, and chi squared. Alongside the images, there were also categories or classes that we wanted our images to be filed into. Two specific classes used, resolved photons labeled “0” and merged photons labeled “1”.

Machine Learning

To run the machine learning algorithm, Keras was used as a backbone with tensorflow running as a backend. Data was saved as images and features as mentioned above. These were then processed using a neural network. More specifically, the images were used to train a convolutional neural network (CNN) while features and the output of the CNN were used to train a deep neural network. Convolutional neural networks are very good at classifying images while the deep neural networks are more specialized for feature analysis. [2] One example of a convolutional neural network is its use to classify handwritten numbers from the MNIST dataset. [5] The deep neural network processed input data such as eta and phi as well as weights of filters coming from CNN to better predict the correct classification for each particle event (0 or 1). In order to train the neural network, 400 events of a variety of classes 0-1 were generated through the particle simulator. Overall the general addition of more layers and complexity to the neural network resulted in better accuracy and prediction power. The flow diagram below illustrates the addition of more data and analysis. See figure 3.

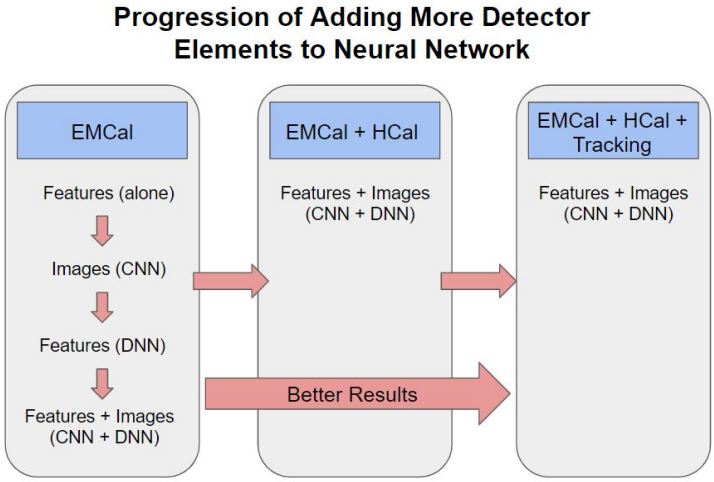

Figure 3) As neural network complexity increased and more data elements were added, the predictive power increased.

Results and Discussion

Data

Data was able to be formatted so that it could be used as input for machine learning successfully. The images of clusters turned out as expected. Merged clusters tended to be larger and cover more area on the 5x5 map than resolved clusters because the energy in resolved clusters was more concentrated and “clean”. This was attributed to the lack of secondary particles hitting nearby any particular cluster. In the case of merged clusters, the reverse was true, a secondary particle had hit close enough to one cluster to trigger a merged cluster, thus the energy signature on the heatmap was wider.

Conventional Methods

As mentioned before, the shower profile from the test beam is used to fit data. The p-value

distribution of the fit for resolved clusters is shown in figure 4. A flat distribution was expected but as seen in figure 4, there is a skew towards lower p-values. For example, applying a cut of 2% should result in a loss of 2% of data samples if the distribution was uniform. However, at a 2% cut, 9.2% (±1.5%) of resolved clusters are lost. Hence there is room for improvement with the use of machine learning.

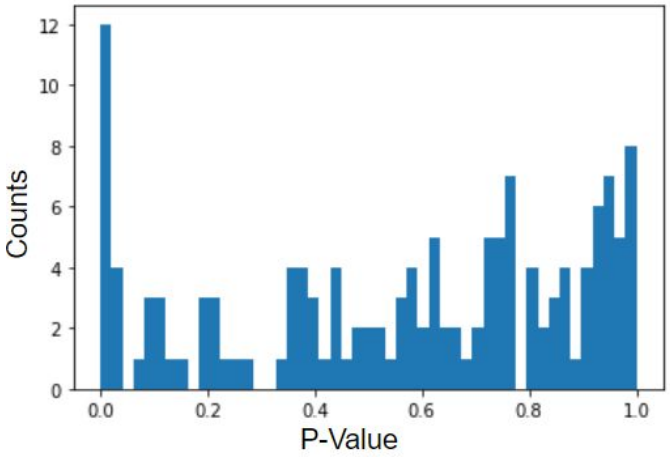

Figure 4) P-value distribution for resolved clusters

Machine Learning

It was able to be shown that as a proof of concept, the purity and accuracy of particle detections can be improved through the application of machine learning. Using this machine learning technique, accuracy values upwards of 98% were achieved. See figure 4. This is a significant improvement over the standard shower shape cutoff accuracy values. This also meant that less data was being thrown away due to a shower shape cut. While some data was still thrown away from an energy cut, this was still an improvement. With more usable data and more accurate predictions, the process of identifying particles and reconstructing jets would proceed much more smoothly and easily.

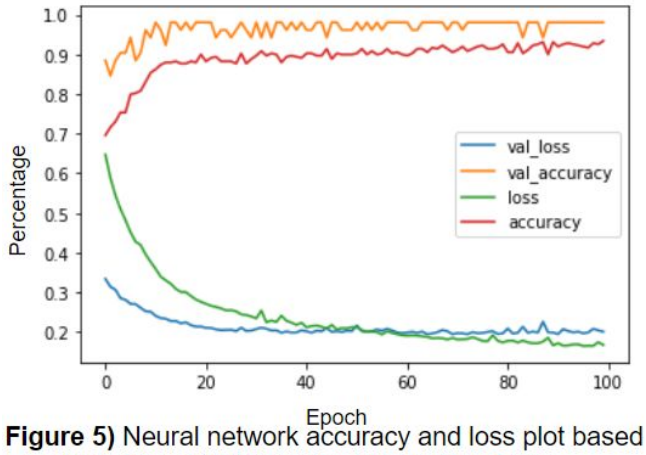

Figure 5) Neural network accuracy and loss plot based on a validation of 50 samples of test data. Here val_accuracy is based on validation data and accuracy is based on training data.Value accuracy reached upwards of 98%.

Advantages

An example of why this machine learning approach is advantageous involves a sample cluster from the study. The cluster in figure 5 shows an event in which the Chi Squared was very low, 0.02394 to be exact. Chi Squared in this case represents the probability that a cluster could be correctly identified. 0.02394 is very low and so any probability cut above 3% used to filter good data samples would likely disregard this cluster and throw it away. However, the machine learning algorithm was able to correctly classify the image as a resolved cluster. The algorithm prediction probability was [0.896273971, 0.103726044], corresponding to [probability resolved, probability merged] respectively. Therefore instead of having to throw data away, it can be salvaged and used by a machine learning algorithm.

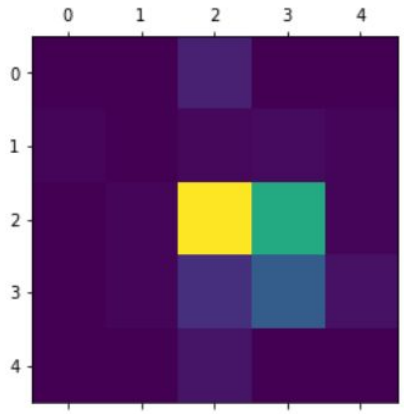

Figure 6) Example of a resolved image that had a Chi Squared of 0.02394 and neural net prediction values of [0.896273971, 0.103726044], class 0 and 1 respectively.

Errors

The machine learning algorithm in this study used only a total of 400 data samples in the form of particle events. 350 were used for training and the remaining 50 were used for validation. In the realm of machine learning these are very small data samples, generally samples on the order of thousands or tens of thousands are used. This affected the testing and validation percentages because each classification carried a lot more weight. Therefore even due to statistical variation the model predictive power varied by several percent. This low number of samples was due to the difficulty of running massive amounts of particle simulations to generate all the data needed. If this process was optimized, then a much larger sample of data would benefit the accuracy and predictive power of the neural network.

Future Work

For future work on this research, we plan to expand the software to use data from additional detectors such as the hadronic calorimeter and tracking detector. The proposition for this is illustrated in figure 1. This would greatly expand the capabilities of the neural network through the addition of more channels through which to process data. Processing data from the hadronic calorimeter would provide information about charged particles such as charged pions. Information from the tracking detector would not only be able to determine the momentum of a particle, but also verify the paths that particles travelled through the detector, thus making it easier to distinguish between merged and resolved clusters as well. All this data combined would make for an information dense neural network with higher predictive power than the current model using only the electromagnetic calorimeter and single channel neural nets. The data format for a new model such as this would consist of 5 classes of images: π^0, π^+, π^-, e^-, and γ (pions, electrons, and photons).